



Universidad Autónoma  
de Madrid

**Biblos-e Archivo**  
Repositorio Institucional UAM

**Repositorio Institucional de la Universidad Autónoma de Madrid**

<https://repositorio.uam.es>

Esta es la **versión de autor** del artículo publicado en:  
This is an **author produced version** of a paper published in:

Machine Learning 112 (2023):921-937

**DOI:** <https://doi.org/10.1007/s10994-021-05999-4>

**Copyright:** © The Author(s), under exclusive licence to Springer Science  
+Business Media LLC, part of Springer Nature 2021

El acceso a la versión del editor puede requerir la suscripción del recurso

Access to the published version may require subscription

*"This version of the article has been accepted for publication, after peer review (when applicable) and is subject to Springer Nature's AM terms of use, but is not the Version of Record and does not reflect post-acceptance improvements, or any corrections. The Version of Record is available online at: <https://doi.org/10.1007/s10994-021-05999-4>"*

# Inference over Radiative Transfer Models using Variational and Expectation Maximization Methods

Daniel Heestermans Svendsen ·  
Daniel Hernández-Lobato · Luca  
Martino · Valero Laparra · Álvaro  
Moreno-Martínez · Gustau Camps-Valls

2021 Springer. Published in Machine Learning (2021). DOI: 10.1007/s10994-021-05999-4

**Abstract** Earth observation from satellites offers the possibility to monitor our planet with unprecedented accuracy. Radiative transfer models (RTMs) encode the energy transfer through the atmosphere, and are used to model and understand the Earth system, as well as to estimate the parameters that describe the status of the Earth from satellite observations by inverse modeling. However, performing inference over such simulators is a challenging problem. RTMs are nonlinear, non-differentiable and computationally costly codes, which adds a high level of difficulty in inference. In this paper, we introduce two computational techniques to infer not only point estimates of biophysical parameters but also their joint distribution. One of them is based on a variational autoencoder approach and the second one is based on a Monte Carlo Expectation Maximization (MCEM) scheme. We compare and discuss benefits and drawbacks of each approach. We also provide numerical comparisons in synthetic simulations and the real PROSAIL model, a popular RTM that combines land vegetation leaf and canopy modeling. We analyze the performance

---

Research partly funded by the ERC under the ERC-CoG-2014 project (grant agreement 647423) and the ERC-SyG-2019 USMILE project (grant agreement 855187), the Ministerio de Ciencia e Innovación project PID2019-106827GB-I00/AEI/10.13039/501100011033, and the projects TEC2016-77741-R, DPI2017-89867-C2-2-R, RTI2018-096765-A-I00, PID2019-109026RB-I00 and PID2019-105032GB-I00.

---

✉ Daniel Heestermans Svendsen  
E-mail: daniel.svendsen@uv.es  
Image Processing Laboratory (IPL), Universitat de València, Spain

Daniel Hernández-Lobato  
Dep. Computer Science, Universidad Autónoma de Madrid, Spain

Luca Martino  
Dep. Signal Theory and Telecommunications, Universidad Rey Juan Carlos de Madrid

Valero Laparra · Álvaro Moreno · Gustau Camps-Valls  
Image Processing Laboratory (IPL), Universitat de València, Spain

of the two approaches for modeling and inferring the distribution of three key biophysical parameters for quantifying the terrestrial biosphere.

**Keywords** Variational Autoencoder · Expectation Maximization · Radiative Transfer Model · Inverse modeling · Density estimation

## 1 Introduction

In many areas of science and engineering, systems are analyzed by running computer code simulations, which act as convenient approximations to reality. Depending on the body of literature, they are known as physics-based, processed-oriented and mechanistic models, or simply just *simulators* [1, 2]. Simulators are ubiquitous in physics, brain, social, Earth and climate sciences [3–5]. Model simulations are needed to understand system behaviour, but also to perform counterfactual studies.

In Earth sciences the use of simulators is of paramount importance. Earth observation (EO) from airborne and satellite remote sensing platforms along with in-situ observations play a fundamental role in monitoring our planet [6–8]. Remote sensing simulators of the involved processes are known as *radiative transfer models* (RTMs). These models describe the complex interactions of scattering and absorption of radiation with the constituents of the atmosphere, water, vegetation and soils. RTMs are useful because they allow us to translate (map) a set of parameter<sup>1</sup> values describing the state of soil, leaf, canopy and atmosphere to at-sensor reflectance or radiance. Such simulations allow for modeling, understanding, and predicting parameters related to the state of the land cover, water bodies and atmosphere.

While modeling and characterizing the involved processes is key, in practice one is typically interested in solving the so-called *inverse problem*; that is, for example, inferring the set of atmospheric or canopy biophysical properties, so that the computed reflectances best fit the remotely sensed ones [9–11]. The problem of inverting the forward model is in general highly *ill-posed* [12]: Different sets of parameter values can map into the same reflectance, thus making it difficult to recover the true set of parameters given a remotely sensed reflectance. This issue has been largely reported in the literature [7, 13, 14] and is, together with the complexity and computational cost of the RTMs as well as the scarcity of labeled data, the main reasons why the inverse problem is a difficult and unresolved one.

Many methods have been proposed for model inversion. Early approaches considered minimizing the (e.g. least squares) error between observations and model simulations stored in big Look-Up-Tables (LUTs). Comparing each observed spectrum with all spectra stored in the LUT proved impractical, leading

---

<sup>1</sup> Note that both the field of remote sensing parameter retrieval and the field of Bayesian inference make heavy use of the word *parameter*. In the former case it refers to the physical parameters over which we are inferring distributions in this work, and in the latter case it refers to distribution parameters (mean, covariance, etc.). The meaning is always clear from context, however.

to gradient-descent techniques in combination with emulators to be proposed in the literature [15]. More advanced approaches have recently exploited machine learning regression algorithms, such as random forests [16, 17], neural networks [18, 19] and Gaussian processes [5, 10, 20–22] to achieve improved multidimensional interpolation capabilities. Treating the inverse problem purely as a regression problem, however, only leads to point-wise estimates, and not a joint probability distribution of the parameters. We argue in this work that when two or more physical variable configurations result in the same spectrum, a conventional inversion method will perform poorly. Such methods that try to predict the cause from the effect will simply predict something in between the possible causes that gave rise to the effect. This problem corresponds to having a multimodal posterior and can be addressed with one of the proposed probabilistic frameworks in this paper.

A Bayesian formalism is useful in order to (1) generate probability density functions (PDFs) of the parameters, and hence account for all moments and uncertainty on the retrieved parameters [23, 24], (2) to incorporate constraints in the form of a-priori parameter distributions, which are often subject of intense debate in the literature [12, 18, 25], and (3) to overcome the limited potential of iterative steepest-descent optimization procedures to locate globally optimal solutions [26, 27]. Furthermore, deriving a generative model provides a straightforward way to perform outlier detection, by measuring the probability of the observed data under the fitted model.

Given some vector of physical causes  $\mathbf{c}$  (atmospheric or canopy properties), the forward RTM model induces a likelihood function  $p(\mathbf{e}|\mathbf{c})$ , which links the causes with the physical effects  $\mathbf{e}$  (reflectance spectra). In this work, we address a general problem: Learning the distribution of the physical parameters or *causes*, instead of only providing a pointwise estimation of these parameters (by statistical or numerical inversion). Provided a dataset of observed effects  $\mathbf{e}'$ , our goal is twofold: learning the marginal density  $p(\mathbf{c})$  and obtaining an approximation of the conditional distribution  $p(\mathbf{c}|\mathbf{e}')$ , which in a Bayesian setting represents the posterior density of the causes given the effects. Note that  $p(\mathbf{c}|\mathbf{e}')$  also represents a probabilistic inverse model, i.e., given  $\mathbf{e}'$  we can obtain a prediction of the causes  $\mathbf{c}$  and related uncertainty measures. Probabilistic inverse modelling, although not so widely used in Remote Sensing applications, has proven to be a powerful tool, providing more general (and hence potentially more valuable) solutions than point-wise approaches, and can help in better understanding the problem itself [27–29].

Since RTMs are generally complex, non-differentiable (i.e. having non-analytical Jacobian) and computationally costly models, mathematical tractability is typically compromised, especially when the aim is to combine RTMs and Bayesian methods. Here, we propose and compare two different approaches which allows us to infer parameters for a non-differentiable simulator. One approach is based on Monte Carlo Expectation Maximization (MCEM) [30] and the other is based on Variational Autoencoders (VAEs) [31]. We will show that each approach has different pros and cons. While the MCEM approach is mathematically elegant, flexible and has good convergence properties, its

application in practice is computationally demanding. On the other hand the proposal based on a simple version of VAE obtains good results and is fast, yet it is not able to describe multimodal distributions. While possible, its extension for multimodal distributions makes the approach more complicated, reducing the good computational properties (see, *e.g.*, [32]). We illustrate these properties in several toy examples of varying sample sizes and complexity, as well as with the PROSAIL RTM [33]. PROSAIL is the combination of the PROSPECT [34] leaf optical properties model and the SAIL [35] (Scattering by Arbitrary Inclined Leaves) canopy reflectance model. In particular, we compare the approaches for inferring the distribution of three key parameters for quantifying the terrestrial biosphere.

## 2 Proposed methodology

### 2.1 Forward and inverse modeling

Notationally, an RTM  $\mathbf{f}$  operating in *forward mode* generates a multidimensional *reflectance/radiance observation (or effect)*  $\mathbf{e} \in \mathbb{R}^{D_e}$  as observed by the sensor given a multidimensional *parameter state vector (or cause)*  $\mathbf{c} \in \mathbb{R}^{D_c}$ , see Fig. 1. Running forward simulations yields a look-up-table (LUT) of input-output pairs,  $\mathcal{D} = \{(\mathbf{c}_i, \mathbf{e}_i)\}_{i=1}^n$ . Solving the *inverse problem* using machine learning implies learning the function  $\mathbf{g}$  using  $\mathcal{D}$ , to return an estimate  $\mathbf{c}_*$  each time a new satellite observation  $\mathbf{e}_*$  is acquired.

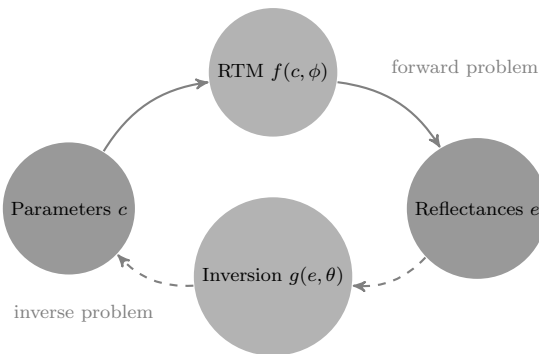


Fig. 1: The forward problem in Earth observation involves taking the system’s structural state as an input, defining representative bio-geo-physical parameters (e.g. vegetation canopy or leaf characteristics), then propagating the solar energy through the atmosphere medium and producing a simulated at-sensor reflectance. The inverse problem involves performing inference over the forward model. We use  $\mathbf{c}$  to denote the model parameters or causes,  $\mathbf{f}$  is an RTM or simulator, and  $\mathbf{e}$  is the simulated output reflectance or effect. Both the forward RTM  $\mathbf{f}$  and the inverse model  $\mathbf{g}$  are nonlinear functions parameterized by  $\theta$  and  $\phi$ , respectively.

## 2.2 Problem setting

Notationally, let us consider then the vector of effects  $\mathbf{e} \in \mathbb{R}^{D_e}$  and vector of causes  $\mathbf{c} \in \mathcal{C} \subseteq \mathbb{R}^{D_c}$ , an RTM model represents the underlying mapping from  $\mathbf{c}$  to  $\mathbf{e}$ , that we denote as  $\mathbf{f}(\mathbf{c}) : \mathbb{R}^{D_c} \rightarrow \mathbb{R}^{D_e}$ . The complete observation model is given by

$$\mathbf{e} = \mathbf{f}(\mathbf{c}) + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{\epsilon}|\mathbf{0}, \sigma^2\mathbf{I}), \quad (1)$$

where  $\mathbf{I}$  is a unit  $D_e \times D_e$  matrix. The observation model defines the likelihood function as

$$p(\mathbf{e}|\mathbf{c}) = \mathcal{N}(\mathbf{e}|\mathbf{f}(\mathbf{c}), \sigma^2\mathbf{I}). \quad (2)$$

Note that by fixing  $\mathbf{c}$ , the conditional probability  $p(\mathbf{e}|\mathbf{c})$  is Gaussian, but as a function of  $\mathbf{c}$  the likelihood is a highly non-linear function due to the dependence on the RTM with the causes, i.e.  $\mathbf{f}(\mathbf{c})$ . We assume an Gaussian prior over  $\mathbf{c}$ 's,

$$p(\mathbf{c}) = \mathcal{N}(\mathbf{c}|\mathbf{m}, \mathbf{S}), \quad (3)$$

where  $\mathbf{m} \in \mathbb{R}^{D_c}$  and the  $D_c \times D_c$  covariance matrix  $\mathbf{S}$  are considered unknown. The posterior density given the observed data  $\mathbf{e}$  over the causes can be expressed as

$$p(\mathbf{c}|\mathbf{e}) \propto p(\mathbf{e}|\mathbf{c})p(\mathbf{c}) = \mathcal{N}(\mathbf{e}|\mathbf{f}(\mathbf{c}), \sigma^2\mathbf{I})\mathcal{N}(\mathbf{c}|\mathbf{m}, \mathbf{S}) \quad (4)$$

Our goals are: (a) learn the prior parameters, vector  $\mathbf{m}$  and matrix  $\mathbf{S}$ , and (b) obtain an approximation of the posterior  $p(\mathbf{c}|\mathbf{e})$ , which serves as an inverse probabilistic mapping from  $\mathbf{e}$  to  $\mathbf{c}$ . We assume that some set of data  $\mathbf{e}$  is given. The two main ways of approaching this problem are a Variational inference (VI<sup>2</sup>) scheme on the one hand, and an expected maximization method on the other. For the VI method we follow the approach of Kingma and Welling [31] and substitute the decoder network with the generative model of Eq. (1). For the MC-based approach we use MC Expectation Maximization [30].

## 2.3 Variational inference method

The idea of variational inference is to optimize the parameters of a *variational posterior* in order to come as close as possible to the *true posterior*. Following [31] we choose a Gaussian variational posterior,

$$q(\mathbf{c}|\mathbf{e}) = \mathcal{N}(\mathbf{c}|\boldsymbol{\mu}_{\text{NN}}(\mathbf{e}), \boldsymbol{\Sigma}_{\text{NN}}(\mathbf{e})), \quad (5)$$

where  $\boldsymbol{\mu}_{\text{NN}}(\mathbf{e})$  and  $\boldsymbol{\Sigma}_{\text{NN}}(\mathbf{e})$  are obtained by tuning a Neural Network (NN) with parameters  $\boldsymbol{\phi}$ . These parameters  $\boldsymbol{\phi}$  are also referred to as the variational

<sup>2</sup> Please note that we use *VI* to abbreviate *variational inference* as in most machine learning literature. This is not to be confused with the abbreviation of *vegetation indices* in the remote sensing literature.

parameters. In order to tune the NN parameters  $\phi$ , we minimize the Kullback-Leibler (KL) divergence between  $q(\mathbf{c}|\mathbf{e})$  and the true posterior  $p(\mathbf{c}|\mathbf{e})$ , i.e.,

$$\begin{aligned} \text{KL} [q(\mathbf{c}|\mathbf{e})||p(\mathbf{c}|\mathbf{e})] &= -\mathbb{E}_{q(\mathbf{c}|\mathbf{e})} \left[ \log \frac{p(\mathbf{c}|\mathbf{e})}{q(\mathbf{c}|\mathbf{e})} \right] \\ &= -\mathbb{E}_{q(\mathbf{c}|\mathbf{e})} \left[ \log \frac{p(\mathbf{c}, \mathbf{e})}{q(\mathbf{c}|\mathbf{e})} - \log p(\mathbf{e}) \right] \\ &= -\mathcal{L} + \log p(\mathbf{e}), \end{aligned} \quad (6)$$

where we have used  $\mathbb{E}_{q(\mathbf{c}|\mathbf{e})}[\log p(\mathbf{e})] = \log p(\mathbf{e})$ . Since  $\log p(\mathbf{e})$  is constant w.r.t. the variational parameters, in order to minimize the KL divergence, we have to maximize

$$\mathcal{L} = \mathbb{E}_{q(\mathbf{c}|\mathbf{e})} \left[ \log \frac{p(\mathbf{c}, \mathbf{e})}{q(\mathbf{c}|\mathbf{e})} \right] = \mathbb{E}_{q(\mathbf{c}|\mathbf{e})} \left[ \log \frac{p(\mathbf{e}|\mathbf{c})p(\mathbf{c})}{q(\mathbf{c}|\mathbf{e})} \right],$$

called the *Evidence Lower Bound* (ELBO). We can split the ELBO into two terms: the first one represents the expected log-likelihood with respect to the variational posterior and the second one is the KL divergence between the variational posterior and the prior, i.e.,

$$\begin{aligned} \mathcal{L} &= \mathbb{E}_{q(\mathbf{c}|\mathbf{e})} \left[ \log \frac{p(\mathbf{e}|\mathbf{c})p(\mathbf{c})}{q(\mathbf{c}|\mathbf{e})} \right], \\ &= \mathbb{E}_{q(\mathbf{c}|\mathbf{e})} [\log p(\mathbf{e}|\mathbf{c})] - \left[ \frac{q(\mathbf{e}|\mathbf{c})}{p(\mathbf{c}|\mathbf{e})} \right], \\ &= \mathbb{E}_{q(\mathbf{c}|\mathbf{e})} [\log p(\mathbf{e}|\mathbf{c})] - \text{KL} [q(\mathbf{c}|\mathbf{e})||p(\mathbf{c})]. \end{aligned} \quad (7)$$

As opposed to the approach in [31] we place a deterministic forward model in stead of a decoder network and fix a low value of noise variance in the likelihood (Eq. 2) in order to reflect the trust in the forward model. This approach is akin to that of [36]. In order to optimize this expression, we perform a Monte Carlo estimation of the expected value (i.e, the first term) [37]. The second term has a simple analytical form as it is the KL divergence between two Gaussians.

Importantly, maximizing  $\mathcal{L}$  with respect to  $\phi$  should make  $\text{KL} [q(\mathbf{c}|\mathbf{e})||p(\mathbf{c}|\mathbf{e})]$  fairly small and hence  $\mathcal{L} \approx \log p(\mathbf{e})$ . The maximization of  $\mathcal{L}$  with respect to the prior parameters,  $\theta = \{\mathbf{m}, \mathbf{S}\}$ , is hence expected to maximize  $\log p(\mathbf{e})$ , which is the maximum likelihood principle for parameter estimation. In practice, we maximize  $\mathcal{L}$  simultaneously with respect to  $\theta$  and  $\phi$ .

The previous approach can be easily extended to the case of having several observed data instances  $\{\mathbf{e}_i\}_{i=1}^N$ . In that case the objective is simply the sum of  $\mathcal{L}_i$ , for  $i = 1, \dots, N$ , where  $\mathcal{L}_i$  is the lower bound corresponding to  $\mathbf{e}_i$ , i.e., the  $i$ -th data instance. This sum can be approximated using mini-batches and optimized using stochastic optimization techniques such as the ADAM algorithm [38]. In this study, we use a mini-batch size of 1 for all experiments. For a proof of convergence of stochastic optimization see [39]. The variational approach is expected to find reasonable values for the prior parameters  $\theta$ , using approximate maximum likelihood estimation, and to provide a recognition model  $q(\mathbf{c}|\mathbf{e})$  that can be used to infer the potential values of  $\mathbf{c}$  given  $\mathbf{e}$ .

## 2.4 Monte Carlo expectation maximization

Another method which can be used to address the learning goals described in Section 2, i.e. to infer the prior parameters from the observed data, and to generate samples from the posterior distribution  $p(\mathbf{c}|\mathbf{e})$ , is the Monte Carlo Expectation Maximization (MCEM) method [30].

We begin by briefly describing the Expectation Maximization (EM) algorithm, which can be used to maximize the likelihood function in models that involve latent variables [40]. This is precisely the scenario considered in Section 2. Namely, given some observed data  $\{\mathbf{e}_i\}_{i=1}^N$ , we would like to maximize

$$\prod_{i=1}^N p(\mathbf{e}_i|\boldsymbol{\theta}) = \prod_{i=1}^N \int_{\mathcal{C}} p(\mathbf{e}_i|\mathbf{c}_i)p(\mathbf{c}_i|\boldsymbol{\theta})d\mathbf{c}_i, \quad (8)$$

as a function of the prior parameters  $\boldsymbol{\theta} = \{\mathbf{m}, \mathbf{S}\}$ . Direct optimization of (8) is intractable, since we cannot marginalize the latent variables  $\mathbf{c}_i$ . The EM algorithm uses the fact that the complete likelihood function  $p(\mathbf{e}_i, \mathbf{c}_i|\boldsymbol{\theta}) = p(\mathbf{e}_i|\mathbf{c}_i)p(\mathbf{c}_i|\boldsymbol{\theta})$  is tractable. Consider the following decomposition of the logarithm of (8)

$$\log \sum_{i=1}^N p(\mathbf{e}_i|\boldsymbol{\theta}) = \sum_{i=1}^N \mathcal{L}(q_i, \boldsymbol{\theta}) + \text{KL}(q_i||p_i), \quad (9)$$

where we have introduced an approximate distribution  $q_i(\mathbf{c}_i)$  and

$$\mathcal{L}(q_i, \boldsymbol{\theta}) = \int_{\mathcal{C}} q_i(\mathbf{c}_i) \log \frac{p(\mathbf{e}_i, \mathbf{c}_i|\boldsymbol{\theta})}{q_i(\mathbf{c}_i)} d\mathbf{c}_i, \quad (10)$$

$$\text{KL}(q_i||p_i) = - \int_{\mathcal{C}} q_i(\mathbf{c}_i) \log \frac{p(\mathbf{c}_i|\mathbf{e}_i, \boldsymbol{\theta})}{q_i(\mathbf{c}_i)} d\mathbf{c}_i. \quad (11)$$

Note that (11) is the Kullback Leibler divergence between  $q_i$  and the exact posterior  $p(\mathbf{c}_i|\mathbf{e}_i, \boldsymbol{\theta})$  for the instance  $\mathbf{e}_i$ .

The EM algorithm maximizes (9) in a two stage iterative process. Assume the current parameter vector is  $\boldsymbol{\theta}^{\text{old}}$ . In the E step, the lower bound  $\sum_{i=1}^N \mathcal{L}(q_i, \boldsymbol{\theta}^{\text{old}})$  is maximized with respect to each  $q_i$  assuming  $\boldsymbol{\theta}^{\text{old}}$  to be fixed. Because  $\sum_{i=1}^N \log p(\mathbf{e}_i|\boldsymbol{\theta})$  does not depend on each  $q_i$ , the solution to this problem consists in setting each  $q_i(\mathbf{c}_i)$  equal to  $p(\mathbf{c}_i|\mathbf{e}_i, \boldsymbol{\theta}^{\text{old}})$ , minimizing  $\text{KL}(q_i||p_i)$  in consequence. In the subsequent M step, each  $q_i(\mathbf{c}_i)$  is held fixed, and  $\sum_{i=1}^N \mathcal{L}(q_i, \boldsymbol{\theta}^{\text{old}})$  is maximized with respect to  $\boldsymbol{\theta}$ , to give new prior parameters  $\boldsymbol{\theta}^{\text{new}}$ . This will cause the lower bound  $\sum_{i=1}^N \mathcal{L}(q_i, \boldsymbol{\theta}^{\text{old}})$  to increase, which will in turn increase the log-likelihood  $\sum_{i=1}^N \log p(\mathbf{e}_i|\boldsymbol{\theta})$ . Critically,  $q_i(\mathbf{c}_i)$  will be computed in this step using  $\boldsymbol{\theta}^{\text{old}}$ , which is fixed. Therefore, the only required integral to evaluate in the M step is

$$\mathcal{L}(q_i, \boldsymbol{\theta}) = \int_{\mathcal{C}} q_i(\mathbf{c}_i) \log p(\mathbf{c}_i|\boldsymbol{\theta})d\mathbf{c}_i + \text{const}. \quad (12)$$



A difficulty, however, is that the posterior  $p(\mathbf{c}_i|\mathbf{e}_i, \boldsymbol{\theta}^{\text{old}})$  is intractable, which makes computing  $q_i$  and hence the integral in (12) challenging. Monte Carlo EM (MCEM), provides a solution to this problem [30]. The intractable integral in (12) is simply approximated by a Monte Carlo average over several samples drawn from  $q_i$ . Namely,

$$\mathcal{L}(q_i, \boldsymbol{\theta}) \approx \frac{1}{S} \sum_{s=1}^S \log p(\mathbf{c}_i^s | \boldsymbol{\theta}) + \text{const.}, \quad (13)$$

where  $\mathbf{c}_i^s$  has been generated from  $q_i$  and  $S$  is the number of generated samples. The convergence properties of MCEM are analyzed in [41].

Recall that the approximate distribution  $q_i$  is targeting the exact posterior  $p(\mathbf{c}_i|\mathbf{e}_i, \boldsymbol{\theta}^{\text{old}})$ . So ideally, we should generate the samples  $\mathbf{c}_i^s$  from the exact posterior. For this, we use Hamilton Monte Carlo (HMC) [42] as in [31]. HMC is a Markov chain Monte Carlo (MCMC) method that can be used to generate (correlated) samples from some target distribution [43]. More specifically, a Markov chain is generated whose stationary distribution coincides with the target distribution. By running the Markov chain for a sufficiently large number of steps one can obtain an approximate independent sample from  $p(\mathbf{c}_i|\mathbf{e}_i, \boldsymbol{\theta}^{\text{old}})$ . HMC has the advantage, when well-tuned, reduces substantially the correlation among samples [43]. For this, it simulates a dynamical system that uses information about the gradient of the posterior, *i.e.*,  $\nabla_{\mathbf{c}_i} \log p(\mathbf{c}_i|\mathbf{e}_i, \boldsymbol{\theta}^{\text{old}})$ , to sample from regions of high posterior probability. In our implementation of MCEM, the HMC procedure consists of 20 leapfrog steps with small step-size (*i.e.*,  $5 \times 10^{-4}$ ) which guarantees that the acceptance rate is high enough. In practice, we only use just one sample to approximate (13). Each time, the Markov chain is initialized at the mode of the posterior distribution, which is found using quasi-newton optimization methods (*i.e.*, L-BFGS). Of course, after optimizing the prior parameters  $\boldsymbol{\theta}$  using MCEM, HMC can be used to generate samples from the approximate posterior distribution  $\hat{p}(\mathbf{c}|\mathbf{e}, \boldsymbol{\theta}) \propto p(\mathbf{e}|\mathbf{c})\hat{p}_{\boldsymbol{\theta}}(\mathbf{c})$ .

## 2.5 Important considerations

Note that both, the variational and MCEM methods, provide an estimation of the parameters  $\boldsymbol{\theta}$  of the prior. Thus, we obtain a Gaussian approximation of the prior, which is denoted here as  $\hat{p}_{\boldsymbol{\theta}}(\mathbf{c})$ . Therefore, both techniques provide the following posterior approximation

$$\hat{p}(\mathbf{c}|\mathbf{e}, \boldsymbol{\theta}) \propto p(\mathbf{e}|\mathbf{c})\hat{p}_{\boldsymbol{\theta}}(\mathbf{c}). \quad (14)$$

However, the variational algorithm provides another posterior approximation given in Eq.(5), *i.e.*,

$$q(\mathbf{c}|\mathbf{e}) = \mathcal{N}(\mathbf{c}|\boldsymbol{\mu}_{\text{NN}}(\mathbf{e}), \boldsymbol{\Sigma}_{\text{NN}}(\mathbf{e})), \quad (15)$$

which yields an important advantage with respect the previous one: given one  $\mathbf{e}$ , using  $q(\mathbf{c}|\mathbf{e})$  we can easily, and at low computational cost, produce a predictive mean  $\boldsymbol{\mu}_{\text{NN}}(\mathbf{e})$  and covariance  $\boldsymbol{\Sigma}_{\text{NN}}(\mathbf{e})$ . The approximation  $\hat{p}(\mathbf{c}|\mathbf{e}, \boldsymbol{\theta})$  on the other hand would require the use of additional Monte Carlo schemes for obtaining a predictive mean and variance, for each new observation vector  $\mathbf{e}$ . Another advantage of the variational approach is the computational speed compared to the MCEM method. However, one advantage of the MCEM scheme is that it can directly handle more practical scenarios (*e.g.*, problems involving multiple posterior modes, heavy tailed distributions, etc.) leading to better performance in terms of smaller error in the parameter estimation of the prior. The variational approach described here would require a different and more general derivation for addressing these scenarios, see *e.g.* [32]. These features of each method are confirmed by the results obtained in our experiments. Implementations of the two approaches can be found at [github.com/dhsvendsen/rtm\\_vi\\_mcem\\_inference](https://github.com/dhsvendsen/rtm_vi_mcem_inference).

### 3 Experiments

We illustrate the strengths and weaknesses of the two approaches, first by means of informative toy experiments: One that studies the computational efficiency of the respective methods, and another which analyzes their ability to handle forward models leading to multimodal posteriors. Following this, we show how these approaches can be used to perform inference over biophysical parameters using an RTM as the forward model.

#### 3.1 On the computational efficiency

In order to analyze the computational efficiency of the two approaches, we consider a simple forward model

$$\mathbf{f}(\mathbf{c}) = f(c_1, c_2) = [2c_1, 2c_2],$$

for which both approaches converge to the true values of the parameters of the prior. We draw the training data  $\{\mathbf{c}_n\}_{n=1}^N$  from the prior

$$p(\mathbf{c}) = \mathcal{N} \left( \begin{bmatrix} 4 \\ 6 \end{bmatrix}, \begin{bmatrix} 1 & 0.6 \\ 0.6 & 1 \end{bmatrix} \right),$$

and pass it through the nonlinear mapping  $\mathbf{f}$  in order to generate the training data  $\{\mathbf{e}_n\}_{n=1}^N$ . Datasets of several sizes  $N = \{50, 500, 1000, 2000\}$  are used for training the models. The model likelihood noise is in all experiments fixed at a negligible value, with  $\sigma^2 = 10^{-7}$ , in order to reflect the trust in the knowledge encoded in the RTMs.

In Fig. 2 we plot an estimate of the average log marginal likelihood of each method on a test dataset as a function of training time (averaged over 40 repetitions). The marginal likelihood is computed using the estimator described in

Appendix A. Observing the test log-likelihood, which is computed after each epoch, we see that the MCEM method converges after 1 epoch (one iteration of the E and M steps). With a training dataset of 50 points, each epoch of training is sufficiently fast that a parallelized version of the MCEM method (in which each E step is done in parallel) converges faster than the VI method. For the non-parallelized algorithm, this is not the case. For larger datasets, VI converges before the completion of 1 epoch of the MCEM algorithm. Since this is a simple toy problem, larger learning rates can be used in the VI method, leading to earlier convergence (just after 1 epoch) for datasets of 1000 and 2000 points. We can conclude from these experiments that the VI approach (as a consequence of stochastic optimization) has a better scaling properties with respect to the dataset size than MCEM.

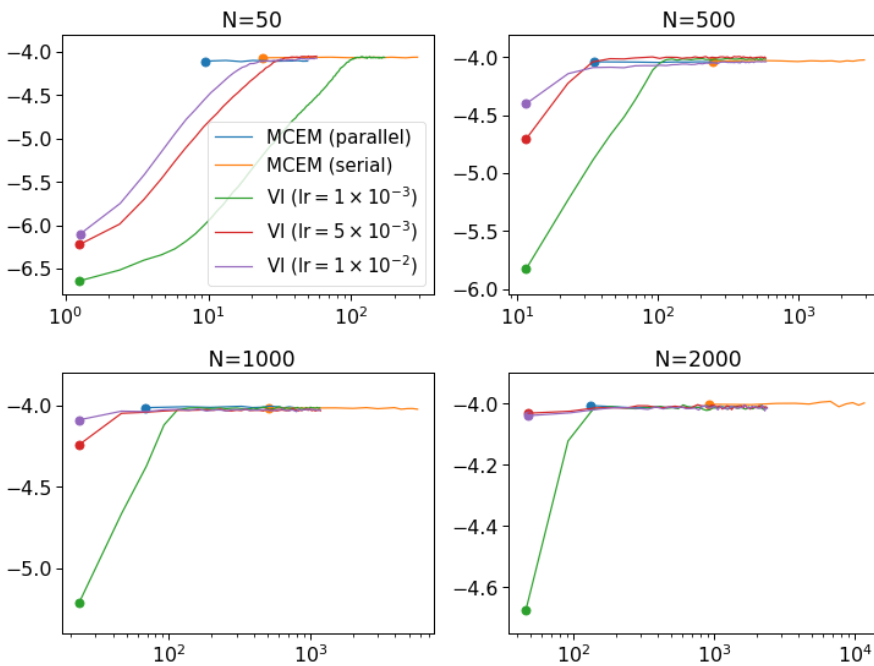


Fig. 2: Marginal log-likelihood of test dataset as a function of training time for different inference methods. Four different sizes of training datasets are used, showing that the VI method is computationally more efficient than the MCEM method for larger datasets.

### 3.2 Dealing with multimodal posteriors

We have seen that when faced with sufficiently large amounts of training data, variational inference performs faster than Monte Carlo sampling methods.

However, since the form of the variational posterior assumed in Eq. (5) is unimodal, we cannot expect it to be able to capture any multimodality in the true posterior. Consider for instance the forward mapping (with  $\mathbf{c} = [c_1, c_2]$ ),

$$\mathbf{f}(\mathbf{c}) = [c_1^2, c_1 c_2].$$

For a given observed  $\mathbf{e} = [e_1, e_2] \geq [0, 0]$  there will always be two possible solutions, namely  $\mathbf{c}^{(1)} = [\sqrt{e_1}, e_2/\sqrt{e_1}]$  and  $\mathbf{c}^{(2)} = [-\sqrt{e_1}, -e_2/\sqrt{e_1}]$  making the posterior inherently multimodal. As stated in the previous sections, we consider  $\epsilon \sim \mathcal{N}(\epsilon|\mathbf{0}, 10^{-7}\mathbf{I})$ . In this example, the prior density is Gaussian with parameters

$$p(\mathbf{c}) = \mathcal{N}\left(\begin{bmatrix} 1 \\ 2 \end{bmatrix}, \begin{bmatrix} 1 & 0.6 \\ 0.6 & 1 \end{bmatrix}\right),$$

from which 500 samples are drawn and passed through  $\mathbf{f}$  to generate the training dataset.

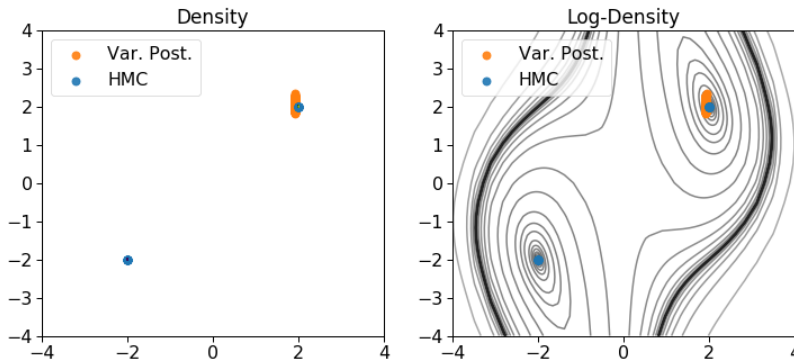


Fig. 3: Contour plots of samples from the true posterior conditioned on the observation  $\mathbf{e} = [4, 4]^\top$ . HMC samples using the prior parameters learned by the MCEM method shown in blue, and samples from the learned variational posterior in orange. The density (left) is so sharply peaked around the two modes that it is more informative to study the log-density (right).

In the process of maximizing the ELBO, the expected log-likelihood with respect to the variational posterior is computed. We can see from Fig. 3, however, that the variational posterior, upon convergence, only captures the positive mode at  $\mathbf{c}' = [2, 2]^\top$  of the true posterior given the observation  $\mathbf{e}' = [4, 4]^\top$ . On the other hand, the MCEM algorithm computes the expected complete log-likelihood with respect to the true posterior as approximated with HMC. As opposed to the variational posterior, HMC does manage to capture both the modes of the true posterior as shown in Fig. 3. The learning algorithm of the MCEM method is therefore more likely to converge to the true parameters of the prior if the posterior is multimodal.

We can see the inability of the variational method to capture the multimodality of the problem from the results of the converged methods given in

Table 1. The fitted parameters of the prior are far from the true ones when compared to the results of the MCEM method which is also reflected in the KL divergence between the fitted and true prior distributions. Multimodality such as this is likely to be observed in the remote sensing experiment latter, as it has been remarked before that different configurations of inputs can lead to the same output making it an *ill-posed inversion problem* [44].

Method	<b>VI</b>	<b>MCEM</b>
Mean	$\begin{bmatrix} 1.20 \\ 1.76 \end{bmatrix}$	$\begin{bmatrix} 1.027 \\ 2.061 \end{bmatrix}$
Covariance	$\begin{bmatrix} 0.619 & 0.682 \\ 0.682 & 1.600 \end{bmatrix}$	$\begin{bmatrix} 1.001 & 0.617 \\ 0.617 & 0.945 \end{bmatrix}$
$D_{KL}$	0.315	0.00581

Table 1: Comparison of methods for inference on a forward model which leads to a bimodal posterior. The first and second columns show the estimated mean vector and covariance matrix of the prior. The third column shows the KL divergence between the fitted and the true prior.

### 3.3 PROSAIL experiment

We now turn to inference in a remote sensing setting using one of the most widely used RTM over the last almost three decades in the field as our physical forward model [45]. PROSAIL is a canopy reflectance model which allows us to relate fundamental vegetation canopy properties, such as, the Leaf Area Index (LAI), and leaf chemical and structural properties, to the scene reflectance for a given set of illumination and sensor (observation) geometry conditions [46]. To perform its simulations, PROSAIL combines two sub-models: PROSPECT [47], which models the optical properties of the leaves; and SAIL [35], which models bidirectional reflectances considering the scattering by arbitrarily inclined canopy leaves in a turbid medium [48]. This combination of models requires the following set of input parameters:

- 1) A set of leaf optical properties (PROSPECT), given by the mesophyll structural parameter (N), leaf chlorophyll (Chl), dry matter (Cm), water (Cw), carotenoid (Car) and brown pigment (Cbr) contents.
- 2) A set of canopy level and geometry characteristics (SAIL), determined by leaf area index (LAI), the average leaf angle inclination (ALA), the hotspot parameter (Hotspot), the solar zenith angle ( $\theta_s$ ), view zenith angle ( $\theta_v$ ), and the relative azimuth angle between both angles ( $\Delta\theta$ ).

We consider PROSAIL for simulating Landsat-8 spectra. This satellite has been widely used in many applications such as cryosphere monitoring, aquatic science and surface water mapping, and vegetation monitoring [49]. Landsat 8’s Operational Land Imager (OLI) includes nine spectral bands with wavelengths ranging from  $0.433\mu m$  to  $1.390\mu m$ , leaving us with an output-dimension of  $D_e = 9$  for our problem. In our experimental setup, we have chosen to work with the most relevant leaf-level parameters to monitor vegetation status and functioning included in PROSAIL, namely Cw, Cm and Chl, resulting in an input dimension of  $D_c = 3$ . The remaining parameters were set constant during our experiments and their values were obtained from previous studies [5] to be representative of realistic cases. Their values can be found in Table 2.

Table 2: Characteristics of the simulations using the PROSAIL model.

<i>Leaf</i>	N	Car	Cbr			
	1.5	8 g/cm <sup>2</sup>	0			
<i>Canopy</i>	ALA	Hotspot	$\theta_s$	$\theta_\nu$	$\Delta\Theta$	LAI
	Spherical	0.01	30°	10°	0	4

Constraining the radiative transfer models with realistic and representative distributions of their inputs is a key part of the RTM inversion process. To facilitate this, in this work we relied on the largest global plant traits database available, the TRY database [50, 51], which contains thousands of leaf data records measured at unprecedented spatial and climatological coverage. Using these data we computed the following empirical mean vector and covariance matrix which was used to sample 2000 values of  $\mathbf{c}$  and pass them through PROSAIL to generate the training data. The empirical mean and covariance (to be compared with the results in Table 3) of the samples are

$$\hat{\mathbf{m}} = \begin{bmatrix} 9.76e^{-3} \\ 1.77e^{-2} \\ 46.2 \end{bmatrix}, \quad \hat{\mathbf{S}} = \begin{bmatrix} 6.42e^{-5} & 5.06e^{-5} & 3.68e^{-2} \\ 5.06e^{-5} & 1.34e^{-4} & -2.86e^{-3} \\ 3.68e^{-2} & -2.86e^{-3} & 288 \end{bmatrix}.$$

The units of the parameters are g/cm<sup>2</sup> for Cm and Cw, and  $\mu\text{g}/\text{cm}^2$  for Chl respectively. Note that the ground truth prior estimated from the TRY database has some probability density in the negative region of parameter space. This is not physically meaningful, but serves the point of illustrating the capabilities of the inference methods. We alter PROSAIL so that it sets every negative parameter to 0 before mapping into spectral space to get a modified likelihood that will lead to more multimodality (since all negative values in  $\mathbf{c}$  will be mapped into the same value, i.e. 0, and then through PROSAIL into a spectrum).

The results of the variational approach to inference over PROSAIL are summarized in Fig. 4. We see that the parameters of the prior are fitted well, which can also be confirmed in Table 3 quantitatively, even though the variational posterior is not able to produce predictive means in the negative domain.

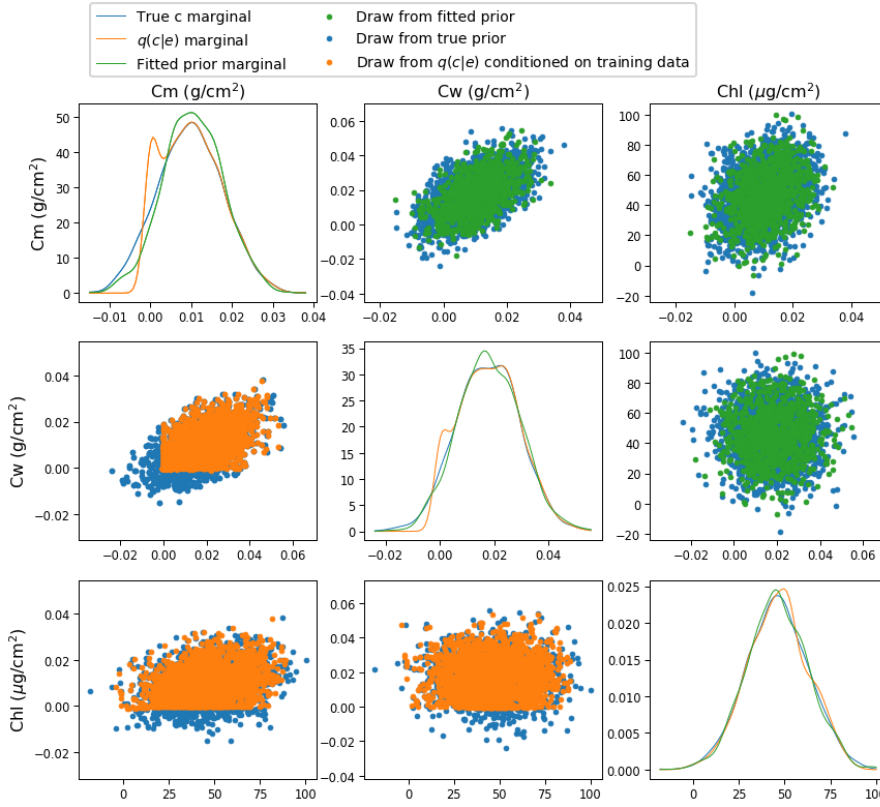


Fig. 4: Results of the variational approach to inference over PROSAIL. The blue points are  $\mathbf{c}$ 's from the training set, while the green points are draws from the fitted prior. The orange points are draws from the variational posterior conditioned on the training  $\mathbf{e}$ 's. The diagonal shows KDE plots of  $\mathbf{c}$  using samples from the ground truth prior (blue), the variational posterior conditioned on training data (orange) and the fitted prior (green).

It is interesting to note that the modification of PROSAIL to truncate negative data, which leads to multimodality, does not prevent the variational approach from estimating the parameters of the prior well.

Nevertheless, the MCEM method is somewhat more accurate than the VI method, obtaining a KL divergence with to the true prior of  $1.23 \times 10^{-2}$  compared to  $2.08 \times 10^{-2}$  obtained using the VI approach. This is to be expected since, as we have seen, the MCEM approach handles multimodality better. We especially foresee a clear difference in results in future work the LAI variable which is difficult to estimate due to its multimodal posterior distribution as pointed out elsewhere [44].

Once the VI method has converged, the neural network which parameterizes the variational posterior can be used as a fast inverse model that maps from observed satellite spectra to biophysical variables. Using the mean outputs that model the mean value of the variational posterior we can obtain

Method	VI	MCEM
Mean	$\begin{bmatrix} 1.02e^{-2} \\ 1.80e^{-2} \\ 45.9 \end{bmatrix}$	$\begin{bmatrix} 1.01e^{-2} \\ 1.81e^{-2} \\ 46.6 \end{bmatrix}$
Covariance	$\begin{bmatrix} 5.32e^{-5} & 4.31e^{-5} & 3.52e^{-2} \\ 4.31e^{-5} & 1.19e^{-4} & 2.18e^{-3} \\ 3.52e^{-2} & 2.18e^{-3} & 280 \end{bmatrix}$	$\begin{bmatrix} 5.49e^{-5} & 4.45e^{-5} & 3.31e^{-2} \\ 4.45e^{-5} & 1.25e^{-4} & -1.74e^{-3} \\ 3.31e^{-2} & -1.74e^{-3} & 292 \end{bmatrix}$
$D_{KL}$	0.0208	0.0123

Table 3: Comparison of methods for inference on biophysical parameters using a radiative transfer forward model. The first and second columns show the mean vector and covariance matrix respectively of the true and the estimated prior over the causes, using  $e = \times 10$  notation for space. The third column shows the Kullback-Leibler divergence between the fitted and the true prior.

good predictive accuracy on a test set as shown in the scatter plots of Fig. 5. Despite the promising results, it is very important to note that we run our experiments using a simplified PROSAIL configuration, keeping some of the input parameters static (see Table 2) and that results can vary greatly in more realistic modeling scenarios.

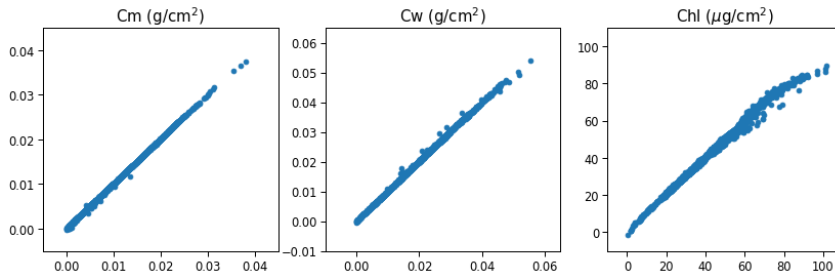


Fig. 5: True values of RTM parameters in test dataset versus mean of variational posterior conditioned on spectra in test dataset. The trained encoder network can thus be used as an effective predictive model

#### 4 Discussion and conclusions

In this work, we approached the long-standing inverse problem in remote sensing of estimating biophysical parameters from observational reflectances. Unlike previous works, we focus on estimating not only the particular parameter point estimates but its full multivariate distribution. We evaluated two different approximations that include an RTM forward model to enforce the inverse estimations to be physically consistent.



Both proposed techniques have different advantages and shortcomings that we illustrated with toy examples and with simulations from the PROSAIL RTM. The MCEM-based approach admits more flexible models while the VAE is computationally more efficient. For instance, while MCEM deals easily with multimodal distributions, this is a challenge for VAE. On the other hand, the convergence time of VAE is orders of magnitude faster depending on the problem. Moreover, the VAE scheme provides a posterior approximation, with a predictive mean and a covariance matrix, implicitly defined by the trained neural network that can be readily evaluated. The experiment involving PROSAIL shows that, while the accuracy of the VAE and MCEM are deemed similar, the computational simplicity of the VAE approach is critical in this problem. Note that including the RTM PROSAIL in the forward-inverse modeling loop increases the time of computation and combining it with MCEM makes it unfeasible especially for large data sets.

We anticipate a wide interest in these techniques for inferring the parameter densities from simulations and then, as further work, from observational satellite data. This will require more accurate and realistic priors; for this we plan to explore mixtures of Gaussians for modeling the prior of the causes as a generalization of the simplified Gaussian model assumed in this work. Likewise, more sophisticated computational methods and variational approaches (e.g., [32, 43, 52]) could be explored in the future.

Finally, we are well aware of the fact that this problem is ubiquitous in other domains of Earth observation and geosciences, and may have implications in climate science too. Inferring parameters is a transversal important topic, not only attached to terrestrial biosphere processes but to the atmosphere, cryosphere and the ocean modeling too. For instance parametrization of small-scale processes such as clouds or biological processes (that are important at the land surface for the exchange of energy and carbon) cannot be explicitly resolved. In this context, learning appropriate parametrizations directly from data may reduce the sources of uncertainties in current models, eventually leading to a deadlock in climate modeling.

## References

1. T.J. Santner, B.J. Williams, and W.I. Notz. *The design and analysis of computer experiments*. Springer Verlag, 2003.
2. B. Wescott. *Every Computer Performance Book*. Create Space Independent Publishing Platform, USA, 1st edition, 2013.
3. M. Raissi, P. Perdikaris, and G.E. Karniadakis. Machine learning of linear differential equations using Gaussian processes. *Journal of Computational Physics*, 348:683–693, Aug 2017.
4. A. Sandberg. Feasibility of whole brain emulation. In *Philosophy and theory of artificial intelligence*, pages 251–264. Springer, 2013.
5. D. H. Svendsen, L. Martino, and G. Camps-Valls. Active emulation of computer codes with Gaussian processes: Application to remote sensing. *Pattern Recognition*, 100:107103, 2020.
6. T. M. Lillesand, R. W. Kiefer, and J. Chipman. *Remote Sensing and Image Interpretation*. John Wiley & Sons, New York, 2008.

7. S. Liang. *Advances in Land Remote Sensing: System, Modeling, Inversion and Applications*. Springer Verlag, Germany, 2008.
8. C. D. Rodgers. *Inverse Methods for Atmospheric Sounding: Theory and Practice*. World Scientific Publishing Co. Ltd., 2000.
9. A. Tarantola. *Inverse problem theory and methods for model parameter estimation*, volume 89. siam, 2005.
10. D. H. Svendsen, L. Martino, M. Campos-Taberner, F. J. García-Haro, and G. Camps-Valls. Joint Gaussian Processes for biophysical parameter retrieval. *IEEE Transactions on Geoscience and Remote Sensing*, 56(3):1718–1727, 2018.
11. R. Zurita-Milla, V. Laurent, and J. van Gijsel. Visualizing the ill-posedness of the inversion of a canopy radiative transfer model: A case study for sentinel-2. *International journal of applied earth observation and geoinformation*, 43:7–18, 2015.
12. B. Combal, F. Baret, M. Weiss, A. Trubuil, D. Mace, A. Pragnere, R. Myneni, Y. Knyazikhin, and L. Wang. Retrieval of canopy biophysical variables from bidirectional reflectance: Using prior information to solve the ill-posed inverse problem. *Remote sensing of environment*, 84(1):1–15, 2003.
13. M. M. Verstraete and B. Pinty. Potential and limitations of information extraction the terrestrial biosphere from satellite remote sensing. *Rem. Sens. Environ.*, 58:201–214, 1996.
14. Y. Knyazikhin. MODIS leaf area index (LAI), and fraction of photosynthetically active radiation absorbed by vegetation FPAR. Technical report, GSFC/NASA, <http://eospso.gsfc.nasa.gov/atbd/modistables.html>, 1999.
15. J. Vicent, L. Alonso, L. Martino, N. Sabater, J. Verrelst, G. Camps-Valls, and J. Moreno. Gradient-based Automatic Look-Up Table Generator for Radiative Transfer Models. *IEEE Transactions on Geoscience and Remote Sensing*, 57(2):1040–1048, 2019.
16. L. Liang, L. Di, L. Zhang, M. Deng, Z. Qin, S. Zhao, and H. Lin. Estimation of crop LAI using hyperspectral vegetation indices and a hybrid inversion method. *Remote Sensing of Environment*, 165:123–134, 2015.
17. M. Campos-Taberner, A. Moreno-Martínez, F. J. García-Haro, G. Camps-Valls, N. P. Robinson, J. Kattge, and S. W. Running. Global estimation of biophysical variables from google earth engine platform. *Remote Sensing*, 10(8):1167, 2018.
18. F. Baret et al. Lai, fapar and fcover cyclopes global products derived from vegetation: Part 1: Principles of the algorithm. *Remote sensing of environment*, 110(3):275–286, 2007.
19. N. Djamai, R. Fernandes, M. Weiss, H. McNairn, and K. Goïta. Validation of the sentinel simplified level 2 product prototype processor (sl2p) for mapping cropland biophysical variables using sentinel-2/msi and landsat-8/oli data. *Remote sensing of environment*, 225:416–430, 2019.
20. G. Camps-Valls, L. Martino, D. H. Svendsen, M. Campos-Taberner, J. Muñoz-Mari, V. Laparra, D. Luengo, and J. Garcia-Haro. Physics-aware Gaussian Processes in remote sensing. *Applied Soft Computing*, 28:69–82, 2018.
21. G. Camps-Valls, J. Runge D. Sejdinovic, and M. Reichstein. A perspective on Gaussian processes for earth observation. *National Science Review*, 6:616–618, July 2019.
22. G. Camps-Valls, J. Verrelst, J. Muñoz Marí, V. Laparra, F. Mateo-Jiménez, and J. Gomez-Dans. A survey on Gaussian processes for earth observation data analysis. *IEEE Geoscience and Remote Sensing Magazine*, (6), June 2016.
23. B. Pinty, I. Andreadakis, M. Clerici, T. Kaminski, M. Taberner, MM Verstraete, N. Gobron, S. Plummer, and J-L Widlowski. Exploiting the modis albedos with the two-stream inversion package (jrc-tip): 1. effective leaf area index, vegetation, and soil properties. *Journal of Geophysical Research: Atmospheres*, 116(D9), 2011.
24. P. Lewis, J. Gómez-Dans, T. Kaminski, Jeffrey Settle, Tristan Quaife, N. Gobron, J. Styles, and M. Berger. An earth observation land data assimilation system (eo-ldas). *Remote Sensing of Environment*, 120:219–235, 2012.
25. C. Atzberger and K. Richter. Spatially constrained inversion of radiative transfer models for improved lai mapping from future sentinel-2 imagery. *Remote Sensing of Environment*, 120:208–218, 2012.
26. Cédric Bacour, Stéphane Jacquemoud, Marc Leroy, Olivier Hautecœur, Marie Weiss, Laurent Prévot, Nadine Bruguier, and Habiba Chauki. Reliability of the estimation of

- vegetation characteristics by inversion of three canopy reflectance models on airborne polder data. *Agronomie*, 22(6):555–565, 2002.
27. Qingyuan Zhang, Xiangming Xiao, Bobby Braswell, Ernst Linder, Fred Baret, and Berrien Moore III. Estimating light absorption by chlorophyll, leaf and canopy in a deciduous broadleaf forest using modis data and a radiative transfer model. *Remote Sensing of Environment*, 99(3):357–371, 2005.
  28. Gabriele Coccia, Amanda L Siemann, Ming Pan, and Eric F Wood. Creating consistent datasets by combining remotely-sensed data and land surface model estimates through bayesian uncertainty post-processing: The case of land surface temperature from hirs. *Remote Sensing of Environment*, 170:290–305, 2015.
  29. Chunfeng Ma, Xin Li, Claudia Notarnicola, Shuguo Wang, and Weizhen Wang. Uncertainty quantification of soil moisture estimations based on a bayesian probabilistic inversion. *IEEE Transactions on Geoscience and Remote Sensing*, 55(6):3194–3207, 2017.
  30. G. C. G. Wei and M. A. Tanner. A Monte Carlo implementation of the EM algorithm and the poor man’s data augmentation algorithms. *Journal of the American statistical Association*, 85:699–704, 1990.
  31. D. P. Kingma and M. Welling. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013.
  32. L. Mescheder, S. Nowozin, and A. Geiger. Adversarial variational Bayes: Unifying variational autoencoders and generative adversarial networks. In *International Conference on Machine Learning*, pages 2391–2400, 2017.
  33. Frédéric Baret, Stéphane Jacquemoud, G Guyot, and C Leprieur. Modeled analysis of the biophysical nature of spectral shifts and comparison with information content of broad bands. *Remote Sensing of Environment*, 41(2-3):133–142, 1992.
  34. Stéphane Jacquemoud and Frédéric Baret. Prospect: A model of leaf optical properties spectra. *Remote sensing of environment*, 34(2):75–91, 1990.
  35. Wouter Verhoef. Light scattering by leaf layers with application to canopy reflectance modeling: The sail model. *Remote sensing of environment*, 16(2):125–141, 1984.
  36. Adam McCarthy, Blanca Rodriguez, and Ana Mincholé. Variational inference over non-differentiable cardiac simulators using bayesian optimization. *arXiv preprint arXiv:1712.03353*, 2017.
  37. Christian Robert and George Casella. *Monte Carlo statistical methods*. Springer Science & Business Media, 2013.
  38. D. P. Kingma and J. Ba. ADAM: a method for stochastic optimization. In *International Conference on Learning Representations*, pages 1–15, 2015.
  39. Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.
  40. A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39:1–22, 1977.
  41. R. C Neath. On convergence properties of the Monte Carlo EM algorithm. In *Advances in modern statistical theory and applications: A festschrift in honor of Morris L. Eaton*, pages 43–62. 2013.
  42. R. M. Neal et al. MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2(11):2, 2011.
  43. L. Martino and V. Elvira. Metropolis sampling. *Wiley StatsRef: Statistics Reference Online*, pages 1–15, 2017.
  44. José Luis Gómez-Dans, Philip Edward Lewis, and Mathias Disney. Efficient emulation of radiative transfer codes using Gaussian processes and application to land surface parameter inferences. *Remote Sensing*, 8(2):119, 2016.
  45. S. Jacquemoud, W. Verhoef, F. Baret, C. Bacour, P.J. Zarco-Tejada, G.P. Asner, C. François, and S.L. Ustin. PROSPECT+ SAIL models: A review of use for vegetation characterization. *Remote sensing of environment*, 113:S56–S66, 2009.
  46. S. Liang. *Quantitative remote sensing of land surfaces*, volume 30. John Wiley & Sons, 2005.
  47. Jean-Baptiste Feret, Christophe François, Gregory P Asner, Anatoly A Gitelson, Roberta E Martin, Luc PR Bidel, Susan L Ustin, Gueric Le Maire, and Stéphane

- Jacquemoud. Prospect-4 and 5: Advances in the leaf optical properties model separating photosynthetic pigments. *Remote sensing of environment*, 112(6):3030–3043, 2008.
48. H. Fang, F. Baret, S. Plummer, and G. Schaepman-Strub. An overview of global leaf area index (lai): Methods, products, validation, and applications. *Reviews of Geophysics*, 57(3):739–799, 2019.
  49. M. Wulder et al. Remote sensing of environment: Current status of landsat program, science, and applications. 2019.
  50. Jens Kattge, Sandra Diaz, Sandra Lavorel, I Colin Prentice, Paul Leadley, Gerhard Bönisch, Eric Garnier, Mark Westoby, Peter B Reich, Ian J Wright, et al. Try—a global database of plant traits. *Global change biology*, 17(9):2905–2935, 2011.
  51. Jens Kattge, Gerhard Bönisch, Sandra Díaz, Sandra Lavorel, Iain Colin Prentice, Paul Leadley, Susanne Tautenhahn, Gijsbert DA Werner, Tuomas Aakala, Mehdi Abedi, et al. Try plant trait database—enhanced coverage and open access. *Global change biology*.
  52. M. Bugallo, L. Martino, and J. Corander. Adaptive importance sampling in signal processing. *Digital Signal Processing*, 47:36–49, 2015.
  53. F. Llorente, L. Martino, D. Delgado, and J. Lopez-Santiago. Marginal likelihood computation for model selection and hypothesis testing: an extensive review. *viXra:2001.0052*, pages 1–58, 2019.
  54. L. Martino, V. Elvira, D. Luengo, and J. Corander. Layered adaptive importance sampling. *Statistics and Computing*, 27(3):599–623, 2017.

## A Marginal likelihood Estimation by Reverse Importance Sampling

In order to evaluate the performance of the methods described in the paper, we use an estimator of marginal likelihood on a test-dataset. More precisely, we use the Reverse Importance Sampling (RIS) estimator [53] described below:

1. Sample  $L$  values  $\{\mathbf{c}_l\}_{l=1}^L$  from the posterior with an MCMC-method. We use Hamiltonian Monte Carlo.
2. Fit a density estimator  $q(\mathbf{c})$  to the samples  $\{\mathbf{c}_l\}_{l=1}^L$ . In this work we fit a Gaussian mixture model, doing cross validation in order to find the best number of components.
3. Sample  $M$  new values  $\{\mathbf{c}_m\}_{m=1}^M$  from the posterior to be inserted in the following estimator:

$$p(\mathbf{e}) \simeq \left( \frac{1}{M} \sum_{m=1}^M \frac{q(\mathbf{c}_m)}{p(\mathbf{c}_m)p(\mathbf{e}|\mathbf{c}_m)} \right)^{-1} \quad \text{where } \mathbf{c}_m \sim p(\mathbf{c}|\mathbf{e}).$$

For the proof and more details see [53]. It is important to remark that the function  $q(\mathbf{c})$  must be a valid probability density for which there are several possible choices. If one chooses  $q(\mathbf{c}) = p(\mathbf{c})$ , RIS becomes the so-called harmonic mean estimator (this name is due to the fact that the corresponding estimator is the harmonic mean of the likelihood values). However, it has been shown that this does not lead to a good estimator. It is possible to show that, in order to ensure finite variance of the resulting estimator, the density  $q(\mathbf{c})$  should have equal or lighter tails than the posterior  $p(\mathbf{c}|\mathbf{e})$  (e.g., see first numerical example in [53]). Gaussian mixture approximations and kernel density estimators of  $p(\mathbf{c}|\mathbf{e})$  are suitable choices for  $q(\mathbf{c})$ . Different alternative estimators of the marginal likelihood are possible mixing MCMC and importance sampling schemes (see [53, 54]).